An assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: Results from the WEB-RADR project

Ola Caster[1] (0000-0002-2259-1716), Juergen Dietrich[2], Marie-Laure Kürzinger[3], Magnus Lerch[4], Simon Maskell[5], G. Niklas Norén[1] (0000-0002-4595-230X), Stéphanie Tcherny-Lessenot[3], Benoit Vroman[6], Antoni Wisniewski[7], John van Stekelenborg[8]

1. Uppsala Monitoring Centre, Box 1051, 75140 Uppsala, Sweden.
2. Bayer AG, Muellerstraße 178, 13353 Berlin, Germany
3. Sanofi, Epidemiology and Benefit-Risk Evaluation, 1 avenue Pierre Brossolette, 91385 Chilly-Mazarin cedex, France
4. Lenolution GmbH, Sybelstr. 7, 10629 Berlin, Germany
5. University of Liverpool, Brownlow Hill, Liverpool, L69 3GJ, United Kingdom
6. UCB Pharma, Braine-l'Alleud, Belgium
7. AstraZeneca Global Regulatory Affairs, Riverside, Granta Park, Cambridge, CB21 6GH, United Kingdom
8. Janssen R&D, Horsham PA, 19044, USA

Corresponding author: Ola Caster, ola.caster@who-umc.org, +46-18-656091

Running heading: Social media for statistical signal detection in pharmacovigilance: Results from WEB-RADR

Acknowledgements

The authors are indebted to the national centres who make up the WHO Programme for International Drug Monitoring and contribute reports to VigiBase. However, the opinions and conclusions of this study are not necessarily those of the various centres nor of the WHO. Further, the authors are indebted to the following colleagues, past or present, within the WEB-RADR consortium who provided technical support that enabled the research presented herein: Beatrice Bourdin, Michael Goodman, Rajesh Gosh, Zeshan Iqbal, Kristina Juhlin, Julia Lien, Carrie Pierce, Amy Purrington, Sue Rees, and Harold Rodriguez.

Compliance with ethical standards

The research leading to these results was conducted as part of the WEB-RADR consortium, (http://webradr.eu) which is a public-private partnership coordinated by the Medicines and Healthcare products Regulatory Agency. The WEB-RADR project has received support from the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under Grant Agreement n° 115632, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

Magnus Lerch provided scientific advice and support to Bayer Pharma AG within the WEB-RADR project and has received compensation for his work from Bayer AG. Antoni Wisniewski is a full-time paid employee of AstraZeneca and holds shares in AstraZeneca. Marie-Laure Kürzinger and Stéphanie Tcherny-Lessenot are employees of Sanofi, which is the marketing authorisation holder of some of the products in the study. All other authors have declared no conflicts of interest.

Abstract

Introduction: Social media has been proposed as a possibly useful data source for pharmacovigilance signal detection. This study primarily aimed to evaluate the performance of established statistical signal detection algorithms in Twitter/Facebook for a broad range of drugs and adverse events.

Methods: Performance was assessed using a reference set by Harpaz et al., consisting of 62 FDA labelling changes, and an internal WEB-RADR reference set consisting of 200 validated safety signals. In total 75 drugs were studied. Twitter/Facebook posts were retrieved for the period March 2012 - March 2015, and drugs/events were extracted from the posts. 4.3 million and 2.0 million posts were retrieved for the WEB-RADR and Harpaz drugs, respectively. Individual case reports were extracted from VigiBase for the same period. Disproportionality algorithms and crude post/report counting were applied in Twitter/Facebook and VigiBase. Receiver operating characteristics (ROC) curves were generated, and the relative timing of alerting was analysed.

Results: Across all algorithms, the area under the ROC curve for Twitter/Facebook varied between 0.47 and 0.53 for the WEB-RADR reference set and between 0.48 and 0.53 for the Harpaz reference set. For VigiBase, the ranges were 0.64-0.69 and 0.55-0.69, respectively. In Twitter/Facebook, at best 31 and 4 positive controls were detected prior to their index dates in the WEB-RADR and Harpaz references, respectively. In VigiBase, the corresponding numbers were 66 and 17.

Conclusions: Our results clearly suggest that broad-ranging statistical signal detection in Twitter and Facebook currently performs poorly and cannot be recommended at the expense of other pharmacovigilance activities.

Key points

- Social media has been suggested as a possibly valuable data source for signal detection in pharmacovigilance. This study focussed on the evaluation of disproportionality analysis in combined Twitter and Facebook data.
- A large number of drugs and a breadth of adverse events were considered. Two different reference sets were used to benchmark predictive performance, one based on labelling changes, and one based on validated safety signals.
- Twitter/Facebook data displayed no predictive value for either of the reference sets, which was contrasted by considerably better performance for the conventional pharmacovigilance data source VigiBase. Therefore, broad-ranging statistical safety signal detection in Twitter and Facebook cannot be recommended.

1 Introduction

During the period 2014-2017, the Innovative Medicines Initiative WEB-RADR (WEB- Recognizing Adverse Drug Reactions) project has addressed key research questions relevant to the potential use of social media for pharmacovigilance.

The advent and massive uptake of social media as a communication tool provides opportunities and challenges in many fields, including pharmacovigilance [1, 2]. One relevant question is whether social media may have value as an independent hypothesis-generating tool in pharmacovigilance, to be used in addition to other data sources such as spontaneous reports of adverse events or electronic health records. If indeed valuable, the vast amount of information generated through social media would require a well-defined approach as regards monitoring, reporting, analysing and evaluating potential adverse reactions, signals and other medical insights related to medicines. The underlying assumption in the utilisation of social media for signal detection is that the type of discussions in social media could either be of a different nature (i.e. different experiences with medications) or take place at a different time than spontaneous reports. If either one of these assumptions holds, social media could indeed be used as a general tool for detection of either different adverse drug reactions (ADRs), or earlier detection of ADRs relative to other data sources, specifically spontaneous individual case safety reports (ICSRs).

The research presented in this paper focuses on the use of social media for aggregate statistical signal detection using spontaneous data as a comparator, specifically VigiBase.

Past investigations in the utility of social media for signal detection have been somewhat limited in either the scope of methods, products, and events (e.g. abuse or misuse) [3-6], or focused on the recognition of adverse events in single posts [7, 8]. In contrast, the work presented here aims to present a comprehensive analysis of the use of social media for the detection of safety signals for a wide range of products using statistical methods. Specifically, our primary aim was to evaluate the predictive ability and timeliness of statistical signal detection using disproportionality analysis in broad-coverage social media like Twitter and Facebook. To this end, both validated safety signals and label changes were used for benchmarking. Additional aims were to investigate the potential utility of statistical signal detection in patient fora, and to assess the clinical relevance of Twitter and Facebook posts for signal detection purposes.

2 Data and methods

2.1 Data extraction and aggregation

2.1.1 Social media data extraction

The raw social media data was provided by Epidemico, one of the WEB-RADR partners. All source data was in the form of free text posts originating from publicly available content from Twitter, Facebook, and various patient fora. This data was acquired either directly from the respective social media platforms, or through a third-party vendor. To maximise signal-to-noise ratio and to increase relevance to drug safety, posts were retrieved using a list of search terms referring to medical product names, including brand names, active ingredients, generic terms, and common misspellings.

After acquisition, the data underwent classification, mapping to medical products, and mapping to

MedDRA®, using the Epidemico algorithm described previously [5, 7, 9]. In this process, each post was assigned an Indicator Score between 0 and 1: a score close to 0 means the post contains language that does not resemble an adverse event discussion (usually spam), whereas a score close to 1 indicates that the post more closely resembles language describing an adverse event. The symptom taxonomy contains a list of MedDRA® Preferred Terms (PTs, 2,167 for the analyses in this work), with a set of colloquial phrases (synonyms) that social media users have used to describe each PT.

The medical product taxonomy contains information for drugs, medical devices, vaccines, cosmetics and dietary supplements. For each of these products, the taxonomy includes both a canonical name, search terms, synonyms (including misspellings and slang terms), and active ingredients to enable downstream grouping at the substance level.

2.1.2 Selection of drugs

Data collection from social media was performed for a pre-specified set of drugs, matching the reference sets used for performance analysis (see Section 2.2). In total 75 substances (or substance combinations) were included, originating either from the publicly available reference set by Harpaz *et al*. [10] or from the internally developed WEB-RADR reference set. Products contributing to the latter are presented in Online Resource 1.

2.1.3 Social media foreground data from Facebook and Twitter

Facebook and Twitter data were acquired and the resulting posts were processed as described in Section 2.1.1.

For the Harpaz substances, 2,024,528 posts were collected with a post date between 1st March 2012 and 31st March 2015 (23% Facebook; 77% Twitter). The total number of posts for the WEB-RADR substances was 4,254,896 (35% Facebook; 65% Twitter), collected over the same period.

The number of Twitter/Facebook posts available for analysis of the Harpaz substances with at least one adverse event, and an Indicator Score of 0.4 or higher was 224,884, whereas there were 465,608 such posts available for the WEB-RADR substances, using the same Indicator Score threshold.

Subsets of data were constructed by applying Indicator Score thresholds, as shown in Table 1.

Table 1. Number of Twitter/Facebook posts for different Indicator Score thresholds.

Indicator Score Threshold	Posts (Twitter/FB) on Harpaz substances	Posts (Twitter/FB) on WEB-RADR substances
0.4	224,884	465,608
0.5	128,199	274,554
0.6	39,461	98,677
0.7	19,120	46,121
0.8	10,028	22,785
0.9	5,232	10,757
0.99	2,130	3,606

FB = Facebook

2.1.4 Social media foreground data from patient fora

In addition to Twitter/Facebook, there are more focused social media channels of potential interest for pharmacovigilance. These patient fora are online communities where patients, family members, and providers come together to discuss diseases and treatments, often limited to a very narrow disease area. Patient fora were selected relevant to the WEB-RADR drugs, and investigated separately from Twitter/Facebook to assess their value for pharmacovigilance signal detection.

The procedure of data provision was the same as for Twitter/Facebook (see Section 2.1.3), with the difference that a single Indicator Score threshold of 0.7 was used. In addition, patient fora posts were only obtained for the WEB-RADR substances, and not for the Harpaz drugs. A total number of 42,721 posts on the 38 WEB-RADR substances from 407 patient fora covering the period 1st March 2012 to 31st March 2015 were collected.

2.1.5 Social media background data

In addition to social media foreground posts, additional posts were collected to provide a broader dataset and more robust estimates of background posting rates. These background posts were collected using the same classifier as the foreground posts, without limitation to the product name, as long as the post contained at least one product.

A total of 4,294,658 posts with an Indicator Score of 0.4 or above were collected, primarily from Twitter (3,056,043 posts, 64%) and Facebook (1,718,892, 36%), with a very small percentage of posts coming from patient fora and discussion groups (310 total).

As with the foreground data, multiple sets of background posts were created using Indicator Score thresholds of 0.4, 0.5, etc.

For each analysis, the applicable foreground data was merged with the background data of the same Indicator Score threshold.

2.1.6 VigiBase data

VigiBase, the World Health Organization (WHO) global database of ICSRs [11], was used as a comparator data source against which social media statistical signal detection performance was contrasted. VigiBase is an established repository of adverse event and suspected adverse drug reaction reports with data from 135 countries. As of 11th March 2018, VigiBase contained 16,870,313 reports in total.

A core extraction of reports from the inception of VigiBase up to March 2015 was performed, although no reports from before March 2012 were used in comparative analyses with social media. Reports were taken from a frozen VigiBase version as of 16th October 2015 containing 14,897,935 reports in total. All active reports were included except those where the submitting country was different from the country in which the event occurred, and each report was assigned a receipt date as the date of the most recent follow-up. No exclusion of reports was performed on the basis of type of report, type of reporter, or other related criteria. Only suspect/interacting drugs were considered.

2.1.7 Aggregated datasets

From the core datasets of social media posts and VigiBase reports described above, corresponding aggregated datasets were generated at the product-event combination (PEC) level. This aggregated

data was subsequently used to compute disproportionality metrics from the different data sources. In the first instance, all combinations of the Harpaz drugs and the various medical concepts defined in the Harpaz reference set [10] were considered, as well as all combinations of the WEB-RADR drugs presented in Online Resource 1 and individual MedDRA® PTs. For each PEC in each considered data source, monthly cumulative counts were generated for the following:

- Number of posts/reports on the combination
- Number of posts/reports on the drug
- Number of posts/reports on the event
- Total number of posts/reports

For social media, foreground and background posts were put together to form the equivalent of a traditional database of ICSRs, like VigiBase. As mentioned above, for patient forum posts, a single Indicator Score threshold of 0.7 was used. For Twitter/Facebook posts, seven different Indicator Score thresholds between 0.4 and 0.99 were considered (see Table 1), each generating a different aggregate dataset. For brevity, these will be referred to as 'Social 0.4', 'Social 0.5' and so on.

For the PECs included in the Harpaz reference set (see Section 2.2.1), monthly cumulative counts were generated for the period March 2012 to March 2015, using February 2012 as the baseline. For the PECs derived from the WEB-RADR drugs in Online Resource 1, cumulative counts were available from April 2012 to March 2015. For the latter set of PECs, one version of cumulative VigiBase counts used the start of VigiBase as baseline, and another version used March 2012 as baseline. Only the latter version was used when comparing social media and VigiBase; this was also the version used to determine which PECs qualified for inclusion into the WEB-RADR reference set according to the definitions of positive and negative controls in Section 2.2.2.

2.2 Reference sets

2.2.1 Harpaz reference set

The publicly available reference set by Harpaz *et al*. is based on US FDA labelling changes performed during the year 2013 [10], which coincides temporally with the collected social media data.

The Harpaz reference contains 62 positive controls, i.e. labelling changes, on 55 drugs and 38 events. Each event is defined by a set of MedDRA® PTs, of which some are considered narrow and some broad with respect to the corresponding event. In this study, only narrow terms were included. Each positive control has an index date corresponding to the date of the labelling revision; for the purposes of this study, the month in which that date fell is used as the index date. The reference set also contains 75 negative controls generated by randomly pairing drugs and events occurring among the positive controls, and manually excluding those with a known, i.e. labelled, association between drug and event.

2.2.2 WEB-RADR reference set

For various reasons, the Harpaz reference in isolation was deemed insufficient to reliably assess the value of signal detection in social media. Firstly, the Harpaz reference set is limited in size. Secondly, its included label changes are severely restricted in geography and time. Finally, and most importantly, whereas labelling changes occur very late in the pharmacovigilance process, safety signals usually occur significantly earlier and are more relevant for protecting patient safety, regardless whether they end up on a product label. The construction of a more relevant reference

set therefore focussed on the concept of the "validated safety signal", i.e. a safety signal with some evidence suggestive of a causal drug/event relationship beyond statistical disproportionality. Additionally, there is intrinsic scientific value in using two different and independent reference sets. Therefore, a larger reference set was generated based on proprietary information on the products listed in Online Resource 1. Positive controls in the WEB-RADR reference set were defined thus:

> *A PEC (on MedDRA® PT level) identified by the manufacturer as a validated signal for the first time in the period between 1st May 2012 and 31st March 2015, that had either (i) at least two posts in the Social 0.7 dataset, or (ii) at least two reports in the aggregated VigiBase dataset, by 31st March 2015, and whose adverse event term belonged to the set of 2,167 PTs included in the symptom taxonomy.*

Each positive control was assigned an index date, defined as the month in which it reached the status of a validated signal. The specific Indicator Score threshold 0.7 was chosen on account of being considered a default quality threshold [9].

Negative controls were defined correspondingly in the following way:

> *A PEC (on MedDRA® PT level) not contained in any HLT linked to any positive control or any listed/labelled PT for the product, and that had either (i) at least two posts in the Social 0.7 dataset, or (ii) at least two reports in the aggregated VigiBase dataset, by 31st March 2015, and whose adverse event term belonged to the set of 2,167 PTs included in the symptom taxonomy.*

Each participating manufacturer generated its own set of positive and negative controls for its included products, and each control was anonymised. All data extraction for this reference set was performed in a decentralised manner at the respective manufacturers, and forwarded in anonymised form for aggregate, central analysis.

2.3 Statistical signal detection in social media data

Disproportionality analysis is the state-of-the-art statistical approach to support the detection of drug safety signals in spontaneous reports [12, 13]. It was therefore selected as the investigational method for evaluating the potential of statistical signal detection in social media data. Disproportionality analysis highlights pairs of drugs and adverse event terms (or groups of terms) with higher-than-expected reporting. Such reporting associations do not in themselves qualify as drug safety signals [13, 14], and will here be referred to as *signals of disproportionate reporting* (SDRs).

2.3.1 Disproportionality analysis measures and algorithms

Two common measures of disproportionality were considered in this study: PRR (Proportional Reporting Ratio) [15], and the IC (Information Component) [16, 17]. Each measure can be applied as part of different signal detection algorithms, whose performance may vary [18]. This study considers four commonly used algorithms, one based on IC, and three based on PRR:

- $IC_{025} > 0$
- $PRR > 2$ and $N \geq 3$
- $PRR > 2$ and $N \geq 3$ and $\chi^2 \geq 4$

- $PRR_{025} > 1$ and $N \geq 3$

where IC_{025} is the lower endpoint of a 95% credibility interval for the IC; χ^2 is the (uncorrected) statistic of a χ^2-test; and PRR_{025} is the lower endpoint of a 95% confidence interval for PRR. These algorithms were applied to VigiBase and the various social media data sources retrospectively in monthly intervals. For social media data, the computations refer to numbers of posts rather than reports.

2.4 Performance evaluations

2.4.1 Analyses at the product-event combination level

Statistical signal detection performance was evaluated in social media and VigiBase data in two ways: receiver operating characteristics (ROC) at fixed time points, and time required to detect positive controls as SDRs. Additionally, time to first social media post was measured.

2.4.1.1 Receiver operating characteristics

ROC curves display sensitivity and specificity at all possible thresholds of a classifier algorithm. In this study, sensitivity and specificity were computed for the four disproportionality algorithms in Twitter/Facebook data (Social 0.4-0.7), forum post data, and VigiBase data, using the Harpaz and WEB-RADR reference sets as benchmarks. In addition, the performance of the raw post/report count (denoted *N*) was tested. This is a useful reference point for disproportionality analysis, and may capture potential issues with the reference set [19].

For the Harpaz reference, data from 1st March 2012 and onwards were used. Positive controls were evaluated in the month prior to their respective index dates, i.e. just before they were labelled. Negative controls were evaluated in December 2013, which is the point in time when their lack of association was established.

For the WEB-RADR reference set, two main analyses were performed. The first included VigiBase data only, and served as a validation of the reference itself. Data were collected from the start of VigiBase to the month prior to the respective index dates of the positive controls, and to March 2015 for negative controls. The second analysis, in which social media data were compared to VigiBase data, was intended to be similar in design to the Harpaz analysis. However, this resulted in too short data collection periods for the positive controls, and consequently unreliable results (see Online Resource 2 for details). Instead, the full data collection period between April 2012 and March 2015 was used for all controls, which means that positive controls were evaluated after their index dates.

For the PRR algorithms presented in Section 2.3.1, ROC curves were generated on the basis of the PRR or PRR_{025} value, and any PEC not meeting the auxiliary conditions on *N* or χ^2 was classified as negative. For some PECs, PRR was mathematically undefined, and for some PECs with zero posts or reports, data were missing to compute both PRR and the IC. All such cases were considered negative classifications.

The area under the ROC curve (AUC) is a common measure of overall predictive performance, which is used in this study also. AUC of 0.5 corresponds to random classification.

2.4.1.2 Timeliness analyses

All timeliness analyses concerned positive controls only. Time required to first detection as SDR (if ever) was analysed for all data sources, using both reference sets. Data were collected for

Twitter/Facebook (Social 0.4/0.7), forum posts, and VigiBase from March 2012 for the Harpaz reference, and April 2012 for the WEB-RADR reference set. For each disproportionality algorithm, the month of discovery of each positive control was compared to its index date.

Further analyses were done for the social media datasets using the WEB-RADR reference set only. First, the month of the first post (if any) of each signal was compared to its index date. Given the available study period (April 2012 to March 2015), this provides a conservative bound on the timing of the first potential warning in social media. Finally, it was investigated whether there were any signals whose first Twitter/Facebook post (within the study period) preceded the first spontaneous report in the corresponding manufacturer's internal database, unrestricted in time.

2.4.2 Post-level content analysis

In addition to aggregate analyses, an assessment of "posts-of-interest" was undertaken to further evaluate the potential value of social media for the identification of drug safety issues.

The primary aim was to quantify the strength-of-evidence in social media for positive controls actually detected using aggregate methods in social media. A secondary aim was to measure the quality of the information present. The assessors were selected from the EFPIA companies, and were pharmacovigilance personnel experienced in the assessment of ICSRs.

- For 25 positive controls (i.e. signals) from the WEB-RADR reference set detected in Social 0.4 before their index date (using the $IC_{025} > 0$ algorithm), the full texts of the corresponding Twitter/Facebook posts were inspected by an expert from the EFPIA company manufacturing that product.
- Each post was assessed using survey questions covering four areas:
    - Does the post contain correct drug and event
    - Is the event an actual adverse experience
    - Is there supporting information in the post
    - Does cumulative evidence exist across multiple posts
- In addition, a comparative analysis was performed by contrasting the results for low Indicator Score posts ($0.4 \leq$ Indicator Score < 0.7) to high Indicator Score posts (Indicator Score ≥ 0.7).

This analysis could be considered an assessment of precision and recall of the Epidemico algorithm on a small but independent test set.

3 Results

3.1 Overview of reference sets and aggregated datasets

Table 2 provides an overview of the various reference sets and their respective coverage in the considered datasets. Generally, the coverage in social media is low.

Table 2. Overview information on the considered combinations of reference set and dataset.

Reference set	Positive controls	Negative controls	Dataset[a]	Positive controls N ≥ 1[b]	Positive controls N ≥ 3[b]	Negative controls N ≥ 1[b]	Negative controls N ≥ 3[b]
Harpaz	62	75	VigiBase	41 (66%)	29 (47%)	36 (48%)	24 (32%)
			Social 0.4	13 (21%)	5 (8%)	17 (23%)	8 (11%)
			Social 0.5	8 (13%)	5 (8%)	8 (11%)	2 (3%)
			Social 0.6	3 (5%)	2 (3%)	2 (3%)	2 (3%)
			Social 0.7	3 (5%)	1 (2%)	2 (3%)	2 (3%)
			Social 0.8	3 (5%)	1 (2%)	2 (3%)	2 (3%)
			Social 0.9	3 (5%)	1 (2%)	2 (3%)	2 (3%)
			Social 0.99	3 (5%)	1 (2%)	2 (3%)	2 (3%)
WEB-RADR	200	5332	VigiBase	197 (98%)	180 (90%)	5072 (95%)	3853 (72%)
			Social 0.4	98 (49%)	75 (38%)	2527 (47%)	1879 (35%)
			Social 0.5	85 (42%)	56 (28%)	2294 (43%)	1653 (31%)
			Social 0.6	46 (23%)	26 (13%)	1461 (27%)	879 (16%)
			Social 0.7	42 (21%)	20 (10%)	1345 (25%)	772 (14%)
			Social 0.8	37 (18%)	19 (10%)	1267 (24%)	679 (13%)
			Social 0.9	35 (18%)	17 (8%)	1216 (23%)	624 (12%)
			Social 0.99	34 (17%)	14 (7%)	1176 (22%)	585 (11%)
			Forum posts	61 (30%)	28 (14%)	1657 (31%)	886 (17%)

[a] 'Social 0.X' means social media data from Twitter and Facebook, with a post-level threshold on the Indicator Score of 0.X. For forum posts, an Indicator Score threshold of 0.7 was used. [b] These figures refer to the specific time points at which data were extracted for positive and negative controls for the purposes of ROC analysis.

Table 3 shows the amount of data for each WEB-RADR substance in each of the different datasets. The variability both within and across datasets is considerable; in particular, there are many drugs with very few social media posts.

Table 3. Number of WEB-RADR substance mentionings in Twitter/Facebook and patient fora; and the number of reports in VigiBase.

WEB-RADR substance(s)	Number of Twitter/FB posts[a]	Percent	Number of patient forum posts	Percent	Number of VigiBase reports	Percent
methylphenidate	13248	28.0%	11178	19.8%	24042	3.6%
topiramate	5190	11.0%	4036	7.2%	15889	2.4%
diclofenac	4310	9.1%	1081	1.9%	66782	10.0%
terbinafine	3706	7.8%	1152	2.0%	19983	3.0%
levetiracetam	2927	6.2%	1372	2.4%	14597	2.2%
vardenafil hydrochloride	2753	5.8%	6023	10.7%	5692	0.85%
propofol	2268	4.8%	435	0.77%	14694	2.2%
carbamazepine	1671	3.5%	1191	2.1%	47209	7.1%
insulin glargine	1619	3.4%	2752	4.9%	26830	4.0%
baclofen	1187	2.5%	2740	4.9%	15667	2.4%
zolpidem	1152	2.4%	2417	4.3%	21593	3.2%
clomipramine	950	2.0%	844	1.5%	8423	1.3%
propranolol	830	1.8%	2184	3.9%	13987	2.1%
zolmitriptan	651	1.4%	207	0.37%	2581	0.39%
tamoxifen	597	1.3%	3821	6.8%	14373	2.2%
estradiol	578	1.2%	2084	3.7%	25924	3.9%
clozapine	450	0.95%	485	0.86%	91511	13.7%
ethinylestradiol,gestodene	432	0.91%	16	0.03%	4300	0.65%
filgrastim	427	0.90%	1366	2.4%	7732	1.2%
oxcarbazepine	306	0.65%	758	1.3%	9412	1.4%
fingolimod	291	0.62%	63	0.11%	17806	2.7%
pegfilgrastim	246	0.52%	1063	1.9%	7190	1.1%
metoprolol tartrate	236	0.50%	574	1.0%	26900	4.0%
clopidogrel	178	0.38%	838	1.5%	36138	5.4%

atenolol	168	0.36%	936	1.7%	23272	3.5%
budesonide	154	0.33%	756	1.3%	13245	2.0%
interferon beta-1b	151	0.32%	26	0.05%	16139	2.4%
letrozole	140	0.30%	4786	8.47%	7891	1.2%
dienogest	126	0.27%	73	0.13%	279	0.04%
omalizumab	116	0.25%	126	0.22%	8469	1.3%
denosumab	63	0.13%	829	1.5%	16954	2.5%
teriflunomide	63	0.13%	20	0.04%	2965	0.4%
artemether,lumefantrine	27	0.06%	4	0.01%	667	0.10%
alemtuzumab	23	0.05%	23	0.04%	3255	0.49%
sorafenib	23	0.05%	35	0.06%	13703	2.1%
romplostim	20	0.04%	51	0.09%	5658	0.85%
dronedarone	16	0.03%	36	0.06%	4344	0.65%
ranibizumab	5	0.01%	100	0.18%	10301	1.6%

[a] At Indicator Score threshold of 0.7; FB = Facebook

3.2 ROC analyses

3.2.1 Harpaz reference set

ROC curves for Twitter/Facebook and VigiBase for the Harpaz reference are shown in Figure 1. The overall performance in Twitter/Facebook is poor, with all ROC curves close to the diagonal, i.e. near random classification. This is in concordance with the low figures displayed in Table 2. Performance in VigiBase is better, with a maximum AUC of 0.67 for IC_{025}.

3.2.2 WEB-RADR reference set

The predictive performance of disproportionality analysis for the WEB-RADR reference set in VigiBase, using all historically available data, is depicted in Figure 2. Although performance is not very good, there is above-random discrimination between positive and negative controls. In a sensitivity analysis where only those positive controls later confirmed as ADRs were used (n=70), the AUC for IC_{025} increased from 0.56 in Figure 2 to 0.62. Here, a confirmed ADR was defined as "a safety signal where sufficient evidence exists to suspect a causal relationship between the signal and the drug and that may require a mitigation action". This means that a positive control may only be classified as a confirmed ADR if the validated signal underwent a full evaluation of all available data by the company. In comparable settings, values as high as 0.74 have previously been observed [19], which suggests that the WEB-RADR reference is challenging. Nonetheless, since this reference yields above-random predictive ability in VigiBase even when evaluating positive controls prior to their index dates, it is considered a valid reference set for the purposes of this study.

The results for the social media datasets and VigiBase when restricted to the period between April 2012 and March 2015, and evaluating all controls at the end of this period, are provided in Figure 3. Although data are collected beyond the signalling dates of the positive controls, social media displays no predictive ability. Indeed, results for Twitter/Facebook are very similar to those observed for the Harpaz reference (see Figure 1). For patient forum posts, there is a peculiar pattern for IC_{025} in the right side of the curve. However, this part of the curve corresponds to an algorithm of about $IC_{025} > -10$ with a majority of true positives having zero posts; hence, there is no practical value in this finding. Results for VigiBase are generally good, and in particular better than in Figure 2. This is expected based on the more extended data collection period, and emphasises the relative underperformance of social media data.

Restricting the positive controls to confirmed ADRs only (see above) did not change the results in any way.

3.3 Timeliness analyses

3.3.1 Harpaz reference set

Time to SDR detection for positive controls of the Harpaz reference is summarised in Figure 4, for Twitter/Facebook and VigiBase data. As expected from the ROC analysis, the results for social media are rather poor.

A single PEC, guanfacine/hallucinations, was detected in Social 0.7 prior to its index date. It was captured by all disproportionality algorithms in March 2013, five months prior to its labelling change. (For reference, detection in VigiBase occurred in May 2012.)

Comparing Twitter/Facebook and VigiBase head-to-head, no PEC was detected earlier in Social 0.7 than in VigiBase, with any disproportionality algorithm. In 31 cases the opposite occurred. For Social 0.4, the corresponding numbers were 4 and 29 PECs, respectively.

3.3.2 WEB-RADR reference set

Timeliness of SDR detection in Twitter/Facebook, patient fora, and VigiBase are shown for the positive controls of the WEB-RADR reference set in Figure 5. Performance in Twitter/Facebook relative to VigiBase is similar to that seen for the Harpaz reference. In patient forum posts, there are more PECs detected in total than in Social 0.7; however, detection appears to be more delayed.

In Social 0.7 there were in total five PECs detected strictly before their index dates, with any disproportionality algorithm. The corresponding numbers were 31 and 1 for Social 0.4 and patient forum posts, respectively, while in VigiBase there were 66 such PECs.

It should be noted that this analysis is biased against VigiBase, since the definition of positive controls excludes PTs that were not considered in the social media extraction pipeline. Any such control might however appear in other data sources, including VigiBase.

For the same positive controls, the distribution of time to occurrence of the first social media post is shown in Figure 6. For Twitter/Facebook, the results clearly show that requiring higher quality posts (i.e. higher Indicator Score) implies later occurrence of the first post. This is expected, as the set of posts with a lower Indicator Score threshold also includes all posts with higher scores. Posting in the considered patient forums occurs generally later than in Twitter/Facebook, which agrees with the SDR timeliness analysis.

Comparing the occurrence of the first social media post (within the study period) to the manufacturers' internal databases of spontaneous reports (unrestricted in time), two positive controls appeared earlier in social media. Both had Indicator Scores between 0.4 and 0.5, and are presumably of low quality. The time differences were small: 1.1 and 0.5 months, respectively. It is important to stress that this number is a lower limit, since there might have been posts on other PECs prior to the start of our study period.

3.4 Post-level content analysis

A total of 631 social media posts were inspected, corresponding to 25 positive controls from the WEB-RADR reference set detected as SDRs prior to their signalling date.

3.4.1 Individual posts

The results of the content analysis of individual posts are presented in Table 4.

Table 4. Results of the content analysis of individual posts.

Question	Yes	Strengthen	Neutral	Weaken
Does the post contain the correct drug?	594 (94.1%)			
Does the post contain the correct medical adverse event?	462 (73.2%)			
If the post contains the correct drug and medical event, is the medical event an actual adverse experience	250 (39.6%)			
Does the post relate the medical event to the drug of interest?	199 (79.6%)[a]			
Is there evidence that patient really took drug?	109 (43.6%) [a]			
Is there information on latency?	24 (9.6%) [a]	8 (33.3%)	16 (66.7%)	0 (0%)
Is there a description on course of the adverse event?	49 (19.6%) [a]	11 (22.4%)	36 (73.5%)	2 (4.1%)
Is there any mention/discussion in the post on risk factors (including lifestyle, medical history, comorbidity, indication) and/or co-medication?	33 (13.2%) [a]	3 (9.1%)	22 (66.7%)	8 (24.2%)
Does post contain patient characteristics: age, gender, weight, height?	7 (2.8%) [a]	1 (14.3%)	6 (85.7%)	0 (0%)
Is there any description as to whether/how the event affected the Quality-of-Life of the patient?	29 (11.6%) [a]			

[a] The denominator for this question are the 250 posts containing the correct drug and medical event, and where the medical event was an actual adverse experience.

These results demonstrate that, at least when using a low Indicator Score threshold, little information can be gleaned from the posts themselves that would aid in the medical triage process, i.e. to determine whether the disproportionality alert should be further considered as a potential safety issue.

Inspection of the posts revealed duplication among the retrieved Twitter/Facebook posts. For example, one PEC had five posts available that corresponded to only two unique Tweets. For another PEC, the same Tweet was duplicated five times. This reduces the information available for triage, and also brings up the question whether these PECs should have been identified as SDRs at all. The issue of duplication was not further investigated here, but merits more attention.

3.4.2 Post series assessment (cumulative strength of evidence)

For each of the 25 PECs, the entire series of posts was assessed for strength of evidence, and the results are presented in Table 5.

Table 5. Results of the questions on cumulative strength of Evidence in the assessment of individual posts.

	Yes	Strengthen	Neutral	Weaken
Consistency of pattern of symptoms	4 (16%)	0 (0%)	24 (96%)	1 (4%)
Consistency of time to onset	2 (8%)	0 (0%)	25 (100%)	0 (0%)
Identifiable subgroup at risk	0 (0%)	0 (0%)	25 (100%)	0 (0%)
Conclusion: would the series of posts (i.e. cumulative evidence) strengthen / neutral / weaken the suspicion of a causal association?		3 (12%)	21 (84%)	1 (4%)

For three positive controls, the inspected posts would have strengthened the signal: in two of the signals, some of the retrieved posts contained enough information for causality assessment (time-to-onset and outcome were present and associated the event with the drug); in the third signal, the

large amount of identified posts (70) in itself was considered evidence-strengthening. Of note is that the posts with evidence for causality both had Indicator Scores greater than 0.7.

3.4.4 Sub-analysis by Indicator Score

The results of the post-level assessment stratified by Indicator Score are given in Table 6.

Table 6. Difference in quality and content between posts with Indicator Score less than 0.7 (LT07) and those with an Indicator Score greater than or equal to 0.7 (GE07).

	LT07 subset	GE07 subset
Does the post contain the correct drug?	488/524 (93.1%)	106/107 (99.1%)
Does the post contain the correct medical adverse event?	387/524 (73.9%)	75/107 (70.1%)
If the post contains the correct drug and medical event, is the medical event an actual adverse experience?	178/524 (34.0%)	72/107 (67.3%)

Adverse events were detected with approximately the same accuracy (~70%) in low- and high-quality posts. However, posts with a low Indicator Score only contained an actual adverse experience 34% (178/524) of the time, whereas higher quality posts, while fewer in number, were much more trustworthy in that respect (72/107, i.e. 67% of these posts contained an actual adverse experience).

The proportion of true positive posts retrieved in high-quality posts compared to that retrieved in low-quality posts is 72/178 = 40%. These 178 true positive posts are a subset of all true positive posts (for the 25 PECs identified as signals), and this proportion of 40% may be considered an upper bound on recall (sensitivity) of the algorithm with Indicator Score threshold >=0.7. These results highlight the trade-off between quality and sensitivity: there are many more posts with a lower Indicator Score than a high one (5:1 ratio), but the average information content in the low-quality posts is much less than those with higher quality. The recall is 2.5 times higher in the low-quality posts but the precision is half of the high-quality posts.

4 Discussion and conclusions

This study investigated the potential usefulness of social media as a broad-based standalone data source for statistical signal detection in pharmacovigilance. Our results provide very little evidence in favour of social media in this respect: in neither of two complementary reference sets, containing validated safety signals and label changes, respectively, did standard disproportionality analysis yield any predictive ability in a large dataset of combined Facebook and Twitter posts. In contrast, individual case report data from VigiBase collected during matching time periods performed well. Likewise, very rarely did the first post or the first occurrence of disproportionality precede the actual time point of signalling, whereas in VigiBase this was much more frequent. The same lack of predictive performance was seen in a non-exhaustive sample of posts from patient fora. Finally, manual assessment of Facebook and Twitter posts underlying 25 early signals of disproportionality showed that only 40% of posts contained the correct drug and the correct event as an adverse experience, and for only three of those 25 signals did the posts strengthen the belief in a causal association.

We have identified four main possible explanations for these results.

Firstly, for the majority of our included drugs there seems to be low activity in the social media platforms we have studied. Indeed, the high number of drugs with very low post counts retrieved with the standard data collection pipeline we have employed is remarkable and suggests that there is limited value of social media as a general pharmacovigilance data source.

Secondly, automatic adverse event recognition in individual posts is difficult, and affects any downstream analysis. In our study, over 600 posts were assessed manually, with precision estimated at 40% for a post-level Indicator Score threshold of 0.4. One potential explanation for this low performance may be that the underlying classification algorithm is not optimised for the rare types of events that are of interest in signal detection. Again, we have employed a standard data collection pipeline that is already in use within the pharmaceutical industry.

Thirdly, the selection and design of reference sets has an obvious influence on the results. We used reference sets that matched our aim, which was to investigate general statistical signal detection. The positive results observed for VigiBase clearly suggest that these references were capable of identifying predictive performance. In fact, the WEB-RADR reference set was restricted to events that the underlying data extraction pipeline was able to identify, which would, if anything, introduce bias in favour of social media. At the same time, both our references contain positive controls populated within the existing pharmacovigilance system, which is largely driven by spontaneous reporting. Therefore, any truly novel signal present in social media would incorrectly appear as a false positive in our study. This issue could only be circumvented by conducting a prospective surveillance study in both data sources, which is laborious and difficult to scale, and was beyond our scope and resources.

Finally, our study was restricted to aggregate measures (i.e. disproportionality analysis and plain counting of reports or posts) that were developed for the purpose of analysing spontaneous reporting data. It is conceivable that other methods tailored to the analysis of social media data would have performed better. However, an argument against this possibility is the low amount of data found in social media in the first place.

The major strength of our study is the breadth and size of the two complementary reference sets employed, which also yielded very consistent results. In fact, the number and types of drugs covered in the two references is very broad and allows for generalisability of the conclusions. In addition, a major aspect of the work is the fact that we did not only use labelled events as positive controls, but also safety signals. As discussed earlier, the concept of a safety signal is more encompassing and relevant to pharmacovigilance than a labelling change. The labelling events of the Harpaz reference constitute an interesting case study, but are not truly representative of the actual day-to-day workings of continual detection and assessment of safety signals, many of which do not end up on product labels, but are subject to further monitoring, e.g. in Risk Management Plans. Lastly, the fact that statistical SDRs were complemented by an inspection of individual posts also solidifies the conclusions. In fact, manual inspection and assessment of the underlying content of an SDR should always be performed, if possible, when ascertaining the value of a new pharmacovigilance data source such as social media.

There are several limitations in the current study that need to be acknowledged. Most importantly, the period covered by the study is quite limited, with only three years' worth of posts being analysed. For the WEB-RADR reference, this precluded our intended ROC analysis at the point of

signalling, and forced us to use all available data beyond the signalling dates. Ideally, any follow-on work would use a longer data collection period generally, and particularly prior to the index dates of the positive controls. Also, we covered relatively few patient fora, and the number of posts retrieved was very small compared to Facebook and Twitter. Well-known patient discussion sites such as Patients Like Me and other subscription sites were not covered in this study. We did not make any distinction between different types of patient fora and lumped all of them into one category, which may have resulted in dilution.

Most other work in this area has focused on identifying and optimising recognition of single adverse events from social media [5, 7, 8, 20, 21], while relatively few papers focused on the actual assessment of utility of social media in providing evidence for ADRs relative to traditional data sources [6, 22]. Other studies do focus on the possible uses of social media above and beyond spontaneous reporting systems, but do not provide specific comparisons in performance [9]. The conclusions in this paper point at the limited utility of social media (at least Twitter and Facebook) even as an additive source for strengthening an initial hypothesis, as the quality of most underlying posts is severely lacking. Other studies [4, 23] did establish that in areas of abuse, large volumes of discussion and new information are readily available in social media and provide a depth and richness of content usually not seen in spontaneous reporting systems. This is not inconsistent with the findings in our study, which did not attempt to single out specific areas of interest.

Our findings of low post counts and high proportions of falsely included posts suggest that improved adverse event recognition from social media posts is a priority area for future research, especially if improved algorithms are able to find and correctly identify adverse experiences across the MedDRA spectrum. Further, whereas we used traditional methods for finding SDRs, there may be methods more suitable for social media, for example methods that take into account the likelihood that a social media post does indeed contain an actual adverse medical event (as opposed to, for example, an indication). Clearly, however, this remains to be demonstrated, and is less of a priority until relevant posts can be retrieved with satisfying sensitivity and specificity. As pointed out above, there may be signals specific to social media that were not part of the Harpaz and WEB-RADR reference sets. This could be investigated through prospective monitoring of social media alongside traditional spontaneous data sources.

All this notwithstanding, it is important to point out that for a majority of drugs, there simply does not seem to be much activity in social media. Any future work should therefore focus on either specific drugs and/or specific areas of interest. Finally, whereas the goal of our work was to assess the utility of social media as a first-line signal detection tool across drugs and events, there are other potential applications of social media in pharmacovigilance that have not been explored here. Some examples include signal strengthening, signal validation, and patients' overall perception of benefit-risk balance. Closed patient fora that are designed around the views and experiences of individual patients might be especially suitable for such in-depth analyses. Even for signal strengthening or signal confirmation of signals detected in other sources, however, it is debatable whether social media (at least Twitter and Facebook) would add value based on our experience from inspecting 600 posts, which yielded very little confirmatory information. A potential issue in using individual posts is that companies, under current regulations, would need to report these to regulatory authorities, increasing the burden on the pharmacovigilance personnel.

In conclusion, our study clearly suggests that general social media like Facebook and Twitter are currently not worthwhile to employ for the purpose of broad-ranging statistical signal detection at the expense of other pharmacovigilance activities. Whereas future improvements to adverse event recognition in social media posts in terms of performance and coverage of events may revise this recommendation, social media is not expected to become a first-line signal detection system. It may, however, serve as a useful complement in specific niche areas.

References

1. Zeng D, Chen H, Lusch R, Li SH. Social Media Analytics and Intelligence. IEEE Intelligent Systems. 2010;25:13-6.
2. Edwards IR, Lindquist M. Social Media and Networks in Pharmacovigilance. Drug Saf. 2011;34:267-71.
3. Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. Proceedings of the 2012 international workshop on Smart health and wellbeing (SHB '12). 2012;10.1145/2389707.2389714:33-40.
4. Sarker A, O'Connor K, Ginn R, Scotch M, Smith K, Malone D et al. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. Drug Saf. 2016;39:231-40.
5. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, Van Le H et al. Evaluation of Facebook and Twitter Monitoring to Detect Safety Signals for Medical Products: An Analysis of Recent FDA Safety Alerts. Drug Saf. 2017;40:317-31.
6. Bhattacharya M, Snyder S, Malin M, Truffa MM, Marinic S, Engelmann R et al. Using Social Media Data in Routine Pharmacovigilance: A Pilot Study to Identify Safety Signals and Patient Perspectives. Pharmaceut Med. 2017;31:167-74.
7. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T et al. Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter. Drug Saf. 2014;37:343-50.
8. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc. 2017;24:813-21.
9. Powell GE, Seifert HA, Reblin T, Burstein PJ, Blowers J, Menius JA et al. Social Media Listening for Routine Post-Marketing Safety Surveillance. Drug Saf. 2016;39:443-54.
10. Harpaz R, Odgers D, Gaskin G, DuMouchel W, Winnenburg R, Bodenreider O et al. A time-indexed reference standard of adverse drug reactions. Sci Data. 2014;1:140043.
11. Lindquist M. VigiBase, the WHO global ICSR database system: basic facts. Drug Inform J. 2008;42:409-19.
12. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18:427-36.
13. CIOMS Working Group XIII. Practical Aspects of Signal Detection in Pharmacovigilance. Geneva, Switzerland: CIOMS; 2010.
14. Delamothe T. Reporting Adverse Drug Reactions. BMJ. 1992;304:465-.
15. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10:483-6.
16. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54:315-21.
17. Norén GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Stat Methods Med Res. 2013;22:57-69.
18. Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S et al. Comparison of Statistical Signal Detection Methods Within and Across Spontaneous Reporting Databases. Drug Saf. 2015;38:577-87.
19. Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance. Drug Saf. 2014;37:655-9.
20. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP '10). 2010:117-25.

21. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. Proceedings of the 2012 international workshop on Smart health and wellbeing (SHB '12). 2012;10.1145/2389707.2389713:25-32.
22. Kürzinger ML, Texier N, Schuck S, Faviez C, Deliens T, Zhang L et al. Web-based signal detection using medical forums data in France from 2005-2015. Pharmacoepidemiol Drug Saf. 2016;25:414.
23. Anderson SL, Bell GH, Gilbert M, Davidson EJ, Winter C, Barratt JM et al. Using Social Listening Data to Monitor Misuse and Nonmedical Use of Bupropion: A Content Analysis. JMIR Public Health Surveill. 2017;3:e6.

Figure captions

1. **Fig. 1** ROC curves for the Harpaz reference set, using data from March 2012 up to the month prior to the index dates for positive controls, and up to December 2013 for negative controls. 'Social 0.X' means Twitter/Facebook data with a post-level Indicator Score threshold of 0.X. AUC ranges between 0.55 and 0.67 in VigiBase, and is 0.53 or lower in Twitter/Facebook. The diagonal represents a random classifier
2. **Fig. 2** ROC curves for VigiBase based on the WEB-RADR reference set. All historical data are used, up to the month prior to the index dates for positive controls, and up to March 2015 for negative controls. AUC values range between 0.56 and 0.59. The diagonal represents a random classifier
3. **Fig. 3** ROC curves for the WEB-RADR reference set, using data from April 2012 up to March 2015 for both positive and negative controls. 'Social 0.X' means Twitter/Facebook data with a post-level Indicator Score threshold of 0.X. AUC ranges between 0.64 and 0.69 in VigiBase, and is 0.55 or lower in all social media datasets. The diagonal represents a random classifier. For the common algorithm $IC_{025} > 0$, sensitivity in VigiBase is 0.38 (at specificity 0.83). For patient forum posts, sensitivity is 0.14 (at specificity 0.88); and for Twitter/Facebook, sensitivity is 0.08 or lower
4. **Fig. 4** Time to SDR detection for the positive controls of the Harpaz reference set, relative to their respective index dates. Data was collected from March 2012 and onwards. 'Social 0.X' means Twitter/Facebook data with a post-level Indicator Score threshold of 0.X
5. **Fig. 5** Time to SDR detection for the positive controls in the WEB-RADR reference set, relative to their respective index dates. 'Social 0.X' means Twitter/Facebook data with a post-level Indicator Score threshold of 0.X. Forum posts were extracted with an Indicator Score threshold of 0.7. Data were collected from April 2012 and onwards
6. **Fig. 6** Distribution of time differences between occurrence of first post and index date, for positive controls in the WEB-RADR reference set. Vertical bars indicate medians and diamonds indicate means. Note that only positive controls with at least one post have been included; the sample sizes are given for each dataset separately (with the total number of positive controls being 200). 'Social 0.X' means Twitter/Facebook data with a post-level Indicator Score threshold of 0.X. Forum posts were extracted with an Indicator Score threshold of 0.7. Data were collected from April 2012 and onwards

Online resource captions

1. WEB-RADR drugs
2. Receiver operating characteristics analysis of WEB-RADR reference set with shorter data collection

Figure 1

Click here to download Figure Fig1_CasterWEBRADR_signal_detection.tiff

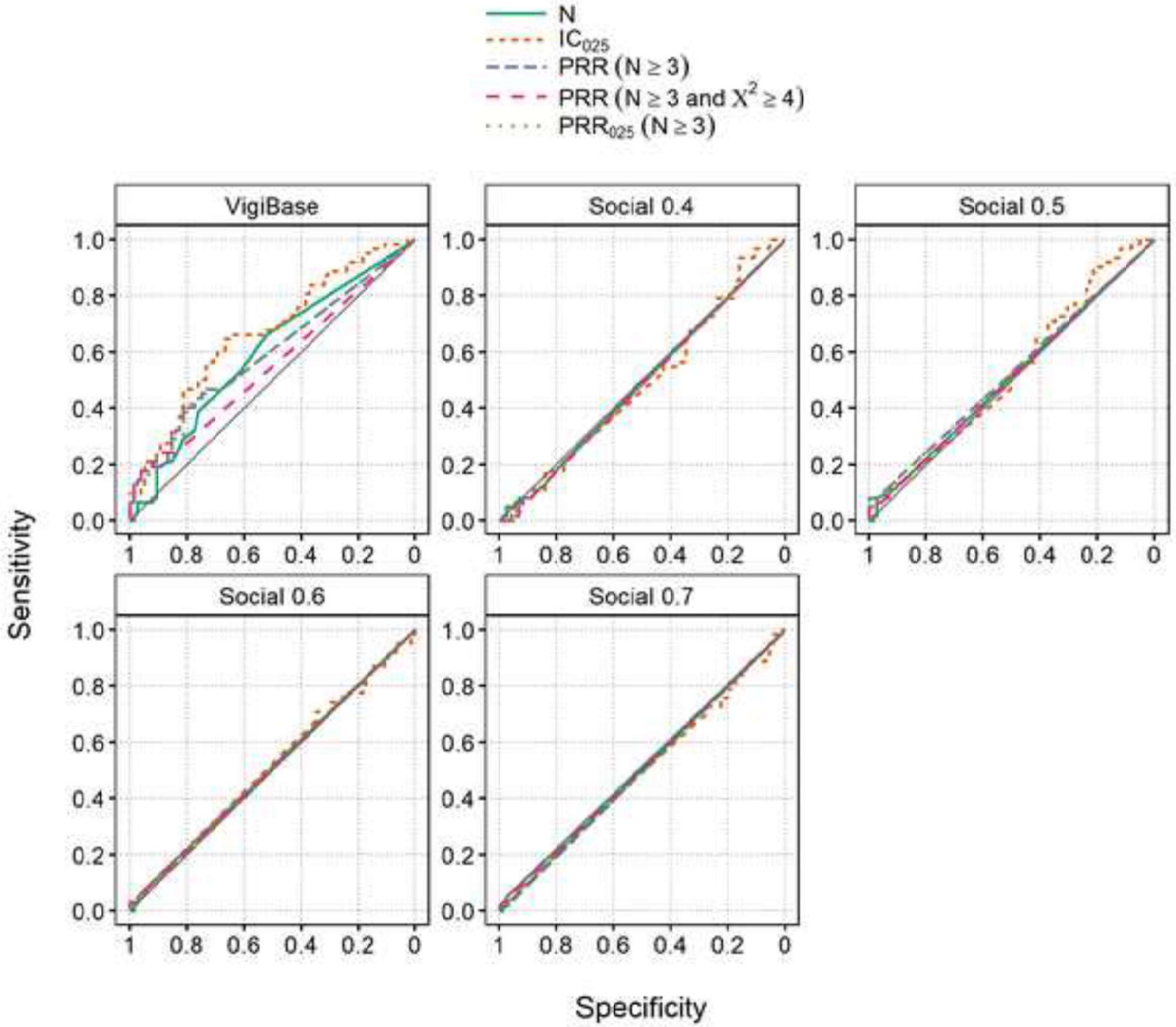

Figure 2

Click here to download Figure Fig2_CasterWEBRADR_signal_detection.tiff

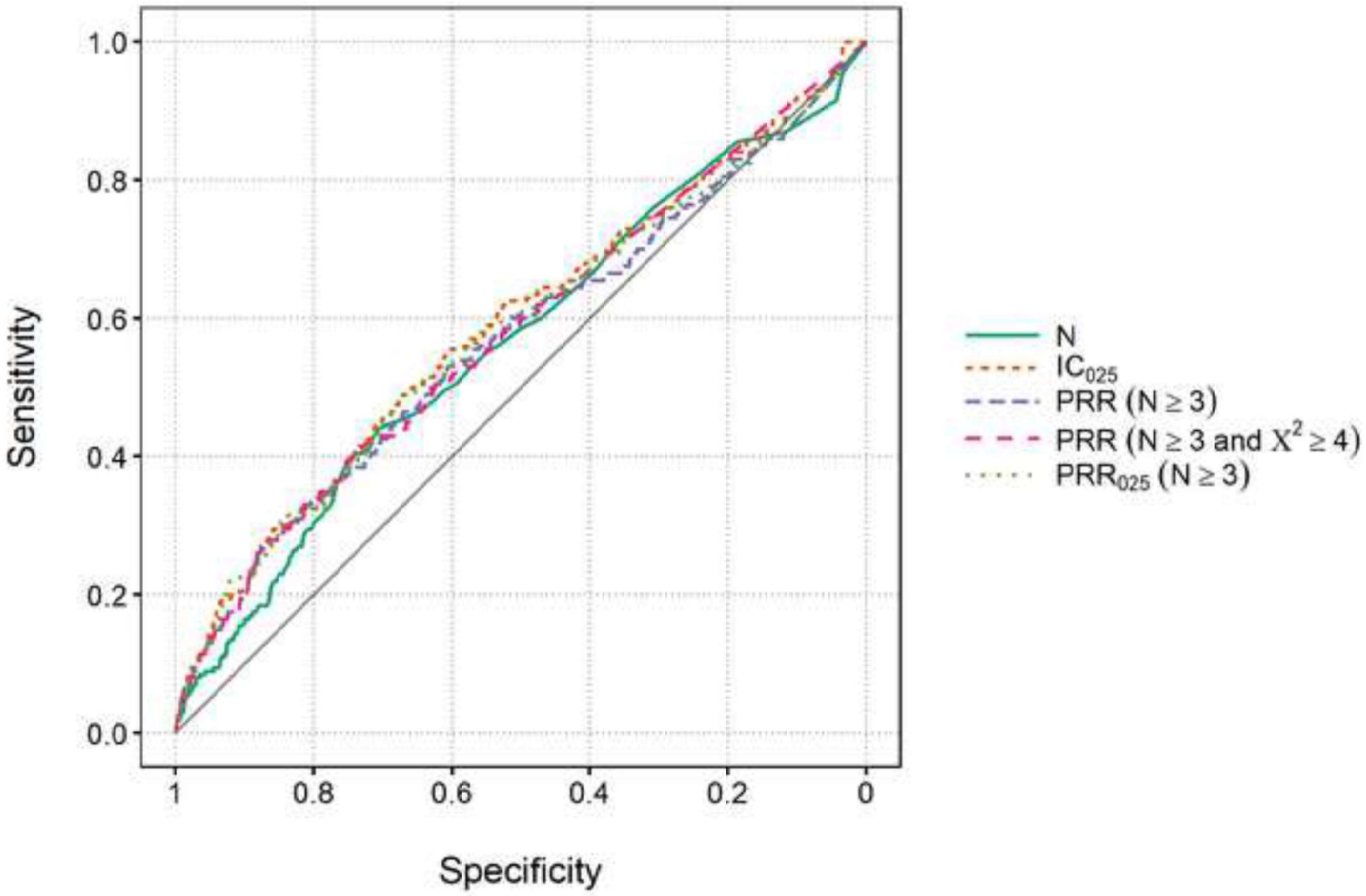

Figure 3

Click here to download Figure Fig3_CasterWEBRADR_signal_detection.tiff

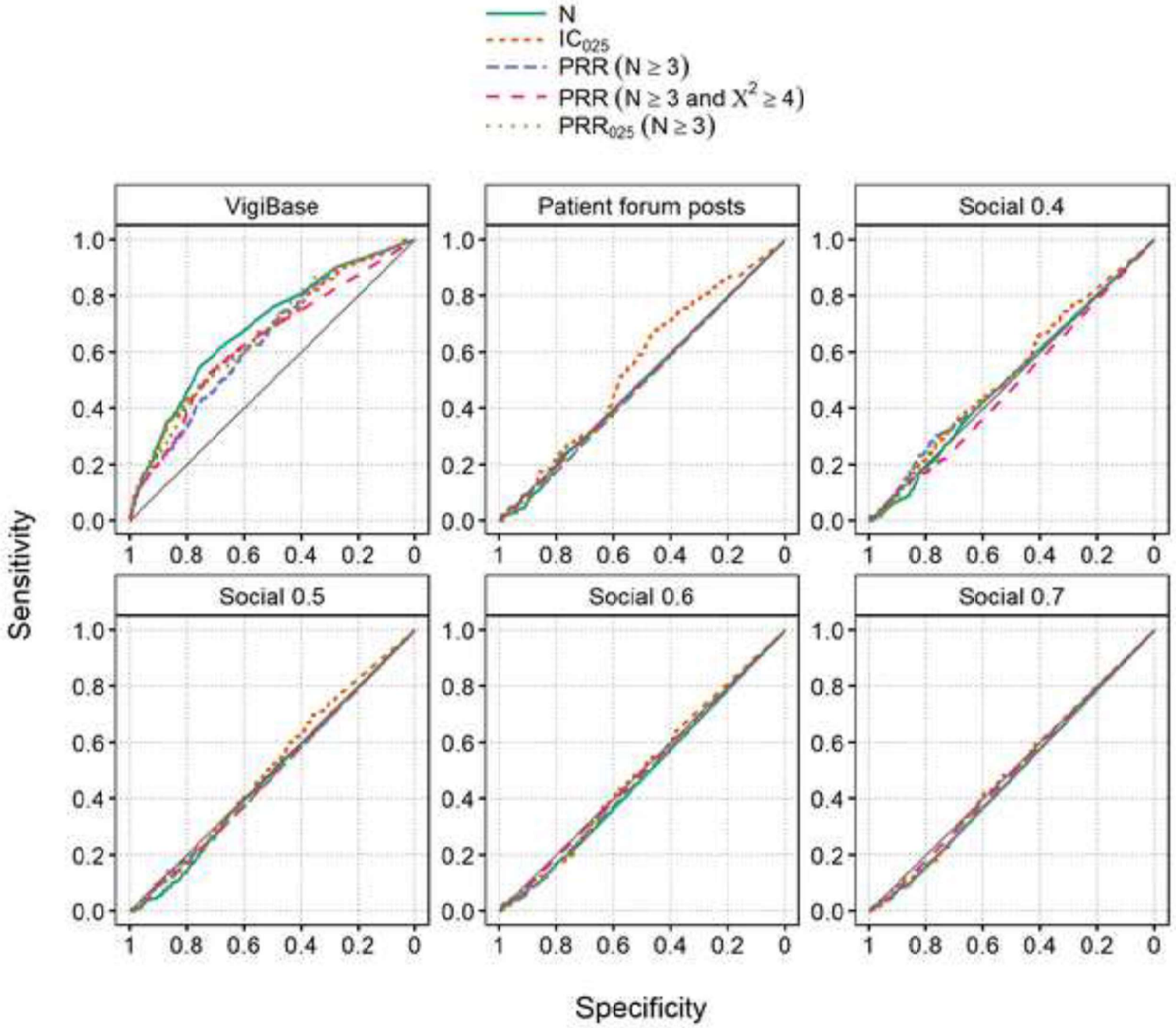

Figure 4

Click here to download Figure Fig4_CasterWEBRADR_signal_detection.tiff

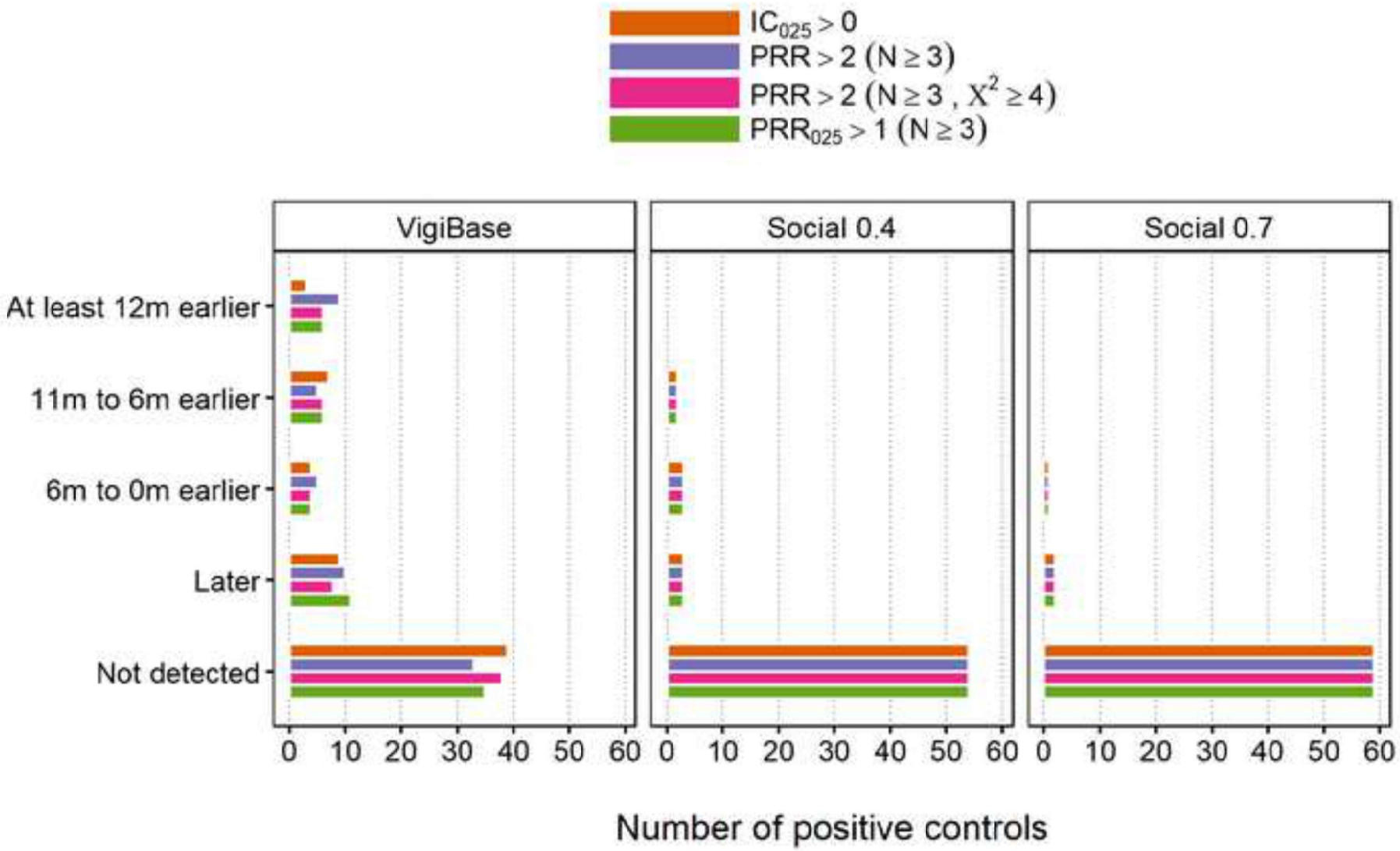

Figure 5

Click here to download Figure Fig5_CasterWEBRADR_signal_detection.tiff

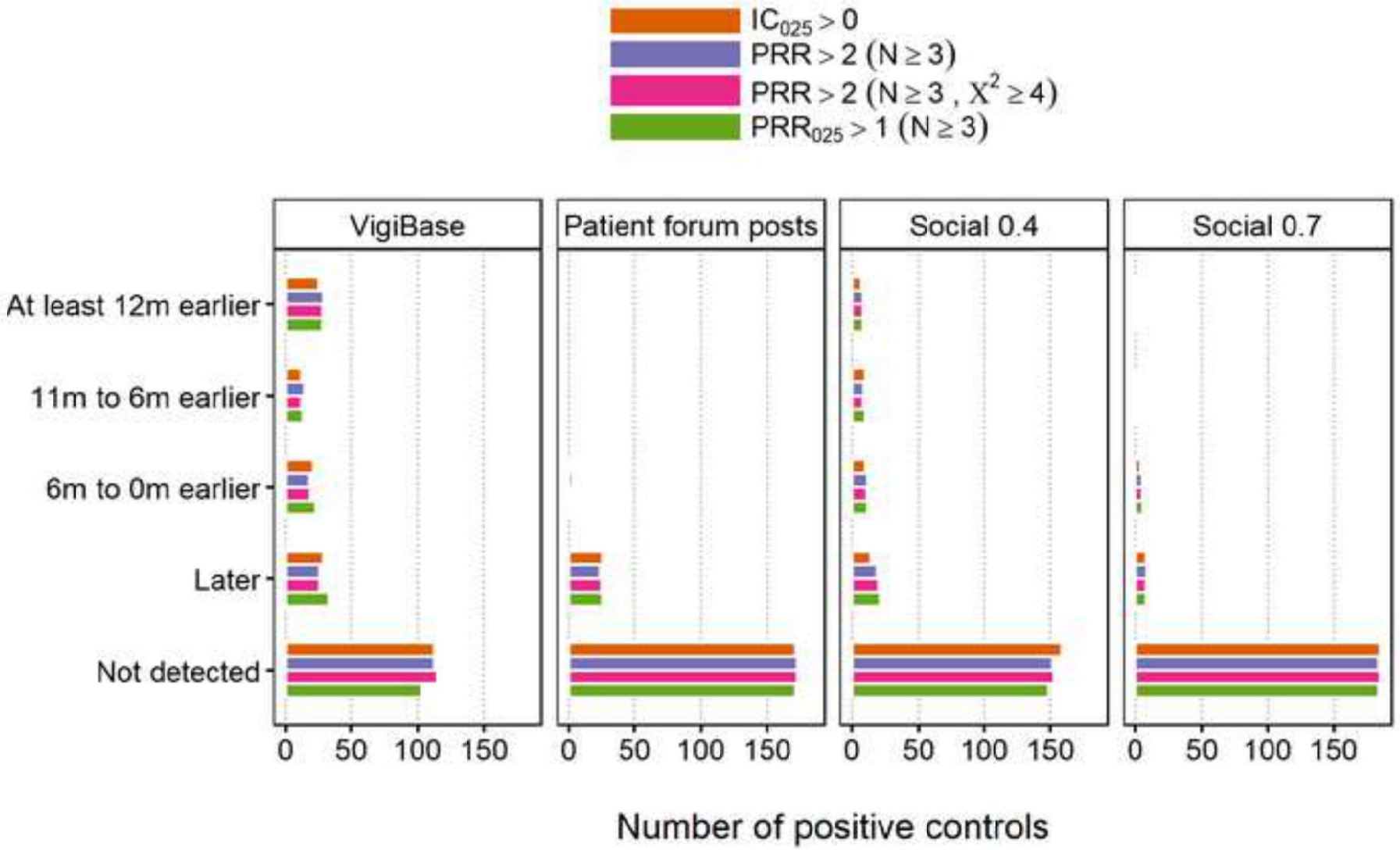

Figure 6

Click here to download Figure Fig6_CasterWEBRADR_signal_detection.tiff